Reverse search in electronic dictionaries

Álvaro Iriarte

1. Introduction

In traditional dictionaries (paper dictionaries, digitized dictionaries, human-readable dictionaries), we can look for information about a word covering the lemmas included in the nomenclature, usually listed alphabetically. After we find the searched word, we can get, in the microstructure, information on:

- meanings of the lemma and their definitions;
- phonetic transcriptions (or respelling);
- grammatical information (morphological, syntactic, semantic, lexical, etc.);
- encyclopaedic or cognitive information, specialised technical field, etc.;
- pragmatic or rhetorical information;
- combinatory possibilities (lexical or syntactic combinations);
- examples and citations;
- etymological information.

But what happens when we don't know the word we're looking for? How can a traditional dictionary help me when I want to know, for example:

- What is the name of 20 units box where are sold beer bottles?
- What euphemism or politically correct term I can use instead of a dysphemism?
- Where do I get information about the verb that combines with the word *passeio* ('walk') to express the sense "to go for a walk"?
- etc.

If the dictionary only presents that information under the entries *grade* ('crate'), the euphemism that you don't know, or the verb *dar* ('give'), I won't be able to find it, because I don't know that *dar* is the word I must use to express that sense ("take a walk" / "to go for a walk"). That's exactly the information I ignore. But that is what happens in most dictionaries: We found the combination *dar um passeio* under the entry ***dar***.

On the other hand, if we register, under the entry ***dar*** a collocation like "dar um passeio", we must also register many other possible combinations:

dar a benção, **dar** a opinião, **dar** a palavra, **dar** a volta, **dar** acordo de si, **dar** alma a, **dar** ares de, **dar** asco, **dar** autorização, **dar** boleia, **dar** cabo de, **dar** carta branca, **dar** certo, **dar** como aberta a (conferência), **dar** conhecimento a, **dar** conta de, **dar** contas de, **dar** corda, **dar** entrada, **dar** entrada em, **dar** erros, **dar** faísca, **dar** faltas, **dar** fé de, **dar** feriado, **dar** forças, **dar** ganas, **dar** instruções, **dar** jeito, **dar** largas a, **dar** licença, **dar** medo, **dar** nas vistas, **dar** o nó, **dar** o sol, **dar** ordem para, **dar** os parabéns, **dar** ouvidos, **dar** patadas, **dar** pena, **dar** permissão, **dar** pontapés, **dar** preguiça, **dar** razão, **dar** saudades, **dar** um abraço, **dar** um beijo, **dar** um conselho, **dar** um golpe de coragem, **dar** um grito, **dar** um nó, **dar** um passo, **dar** um salto, **dar** um salto lá, **dar** um som, **dar** um suspiro, **dar** uma ajuda, **dar** uma bofetada, **dar** uma conferência, **dar** uma corrida, **dar** uma injecção, **dar** uma mão, **dar** uma negativa, **dar** uma ordem, **dar** uma palavrinha, **dar** uma queda, **dar** uma sugestão, **dar** uma vacina, **dar** vontade, **dar** (dois dedos de) conversa, **dar** um filme, não **dar** uma para a caixa, etc.

Which we can also join close to 570 set phrases beginning with the word *dar* (plus 46 initiated with "*não dar*") recorded by Simões (1994) in his *Dicionário de Expressões Populares Portuguesas,* or 570 set phrases recorded by Ramalho (1985) in his *Dicionário Estrutural, Estilístico e Sintáctico da Língua Portuguesa.*

These lexical combinations are unproblematic when we use the dictionary as a tool for decoding. It's not especially hard to discover the meaning of the combination *dar um passeio* if you look it up in a dictionary.

The problem arises in a particular way when we intend to proceed to codification or text production, because the combinatory possibilities vary from one language to another (and, in general, the lexical-syntactic and pragmatic uses).

What happens when I want to express the idea "take a walk" or "to go for a walk" in Portuguese? Let's just say that I do not know the construction *dar um passeio*, but I do know the Portuguese word *passeio* is "walk" in English. Will dictionaries help me?

It is in the entry **passeio** that the user should be able to find the information. A good lexicographic treatment of words implies collecting their different combinatory possibilities. These combinatory possibilities of words are determined as much co-textually (that is, by the linguistic context) as pragmatically and contextually. I mean, the meaning of a word or group of words can be defined or delimited by context, or by other lexical units, which, along with that word, compose the syntagma.

Idiomatic expressions (such as "it's raining cats and dogs") should be collected in the syntagmatic part of the lexicographic article[109], since it is not always possible to associate a multi-word expression to a concrete meaning of the words it is composed of. But, what happens in the case of lexical collocations? How should collocations be collected in dictionaries?

Take another example:

tirar uma conclusão ('draw a conclusion')
tirar uma fotografia ('take a picture')

[109] In the case of "it's raining cats and dogs", in encoding dictionaries, should be collected under the entry **rain** (but I will not talk about it in here).

The way most dictionaries are made, users may only find this information under the entry ***tirar*** (that's just exactly the information they ignore!):

> **tirar** [. . .] v. [. . .] **II.** Como verbo suporte de predicação, combina-se com nomes ... 1. Actos físicos ≃ FAZER. ... *Tirar uma fotografia* ... **2.** Actos morais ≃ OBTER. *Tirar conclusões* (Academia).
>
> **tirar** [. . .] v. **31.** captar (imagem), . . . fotografando, . . . : *t. uma foto de uma pessoa.* (Houaiss)
>
> **tirar** V. t. d. [. . .] **15.** Fazer (uma fotografia [2]); *Fique aí quieto, vou tirar a fotografia.* **16.** Fazer tirar, parar para tirar (uma fotografia [2]): *Aprontou-se toda para tirar o retrato.* [. . .] ([Aurélio]Ferreira).
>
> **TIRAR** v. tr. [. . .] Tirar (alguém) o retrato, fazer-se retratar: *Fui tirar o retrato para a carteira de identidade.* || Tirar o retrato a alguém, fazer-lhe o retrato: [. . .] ([Caldas] Aulete).
>
> **Tirar**, V. t. [. . .] Derivar: *tirar conclusões.* [. . .] ([Cândido] Figueiredo).

Only in the dictionaries *Houaiss* and *Aulete Digital* we have information under the entry ***fotografia*** (but nothing under the entry ***conclusão*)**. In the *Dicionário Houaiss*, in the grammatical information under the entry ***dar***, we can read:

> **a.3**) por sua importância, diversas acepções de *dar,* usado como verbo-suporte, estão registadas no corpo deste verbete; diversas outras devem ser procuradas pelo substantivo que faz parte do objeto direto, como de hábito no restante dicionário. (Houaiss, *s.v.* **dar**).

Another exception is the *Dicionário Básico da Língua Portuguesa,* which has a good resolution of the issue, because we have information about the collocation *tirar uma conclusão* under the entry **conclusão:**

> **tirar**... [...]
>
> **S. 8.** Tirar + nome, sentido VIII, equivale a um verbo simples: CONCLUIR (FRASE 1), [...] (Vilela).
>
> **conclusão**... [...]
>
> [...] // (pessoa) **tirar conclusões:** (5) — *Que conclusões podemos tirar da sua atitude?* • (6) — *Não quero tirar conclusões erradas do caso.* [...]
>
> **S.** • *Tirar conclusões* (frases 5, 6) tem como sins.: CONCLUIR, TIRAR ILACÕES OU DEDUÇÕES ... (Vilela).

2. Reverse search[110]: onomasiological dictionaries and encoding dictionaries

An onomasiological dictionary is the opposite of a traditional dictionary. When we use an ordinary dictionary, we have a word in our head and we want to look up its meaning. When we use an onomasiological dictionary, we have a concept in our head, and we want to look up which word, or words, best expresses it. In these

[110] A reverse dictionary is not the same as a reverse-order dictionary. In a reverse-order dictionary the alphabetical sorting is done from right to left. They let the user browse the dictionary searching by the end of the word, instead of its beginning.

dictionaries we can search an idea or concept in a descriptors structure or a structured list of concepts sorted by subjects (summary tables), together with a list of hypernyms or broader terms (categories, general ideas) that lead the reader to the searched word.

The major disadvantage of this sort of dictionaries is that the organization and classification of information varies from dictionary to dictionary because the knowledge's organisation varies from author to author (Béjoint, 2004: 15).

There is a long tradition of onomasiological dictionaries production for the European languages, especially in the nineteenth and twentieth centuries (see some examples in Sousa, 1995, s.v. diccionario ideológico).

Another type of reverse dictionary is the encoding dictionary. It is important to distinguish between encoding dictionaries and onomasiological dictionaries. In encoding dictionaries you don't look for an idea to find a word. Encoding dictionaries supply you with information about the co-text (that is, about the lexical, syntactic or semantic combinatory capacity) and the context (information of a pragmatic kind). So, encoding dictionaries supply you with information about the co-text and the context of certain lexical choices in order to transmit certain concepts.

In an onomasiological dictionary, you should look for the information in a kind of structure that is organized by subjects through hyperonyms or generics. These allow us to get to the word we are looking for. So, to find a word that means *forte* ('strong') applied to drinks, we should look under the entry **intensidade** ('intensity') or something like that. In onomasiological dictionaries the lemma is this synonym or hyperonym (**intensidade**):

> **intensidade:** forte [bebidas]

In encoding dictionaries the lemma is a word, or set of words, about certain information we are looking for. This is the opposite of what happens in onomasiological dictionaries, in which you look for an idea or a concept from its synonym or hyperonym. In an encoding dictionary, the user should look for the lexeme that expresses the idea of *forte* ('strong') applied, for example, to the word *café* ('coffee') under the entry **café**. In encoding dictionaries you search words, not concepts. You are using the "word" as a unit, as lemma, and you keep the traditional alphabetical order:

> **café** ... [muito intenso] > *forte*
> **ódio** ... [muito intenso] > *mortal, figadal*
> **passeio** ... [realizar] > *dar um passeio*

Onomasiological dictionaries are paradigmatic dictionaries while encoding dictionaries are paradigmatic and syntagmatic dictionaries. The main limitation of the onomasiological dictionaries, and in general of all the so-called paradigmatic dictionaries (onomasiological dictionaries, synonym and antonym dictionaries, and so on) is that they just enumerate synonyms, with no examples, no exact indication on contexts of usage, etc: "...en gran parte, son poco satisfactorios, ya que se limitan a dar una simple enumeración de sinónimos, sin indicaciones exactas sobre

denotación, connotación, situaciones de uso, etc., y, sobre todo, sin ejemplos." (Haensch, 1982: 178).

3. Implementation of reverse search functionalities in electronic dictionaries

The reverse search capabilities transform electronic dictionaries not only in an ideological or conceptual dictionary but also in an encoding dictionary. It allows us to find not only the word that corresponds to an idea (like onomasiological dictionaries), but also which word we can associate to another to express an idea (like encoding dictionaries).

The reverse search capability in electronic dictionaries can help overcome some of the limitations present in onomasiological dictionaries and encoding dictionaries on paper. This feature can be more than a simple search tool. I can imagine the possibilities of reverse research into new lexicographical products, developed with more scientific rigor and with a more systematic microstructure. In this electronic dictionaries it should be possible to search anywhere on the macrostructure and the microstructure.

With a simple search tool in the electronic dictionary, we can find related lexical units (synonyms, quasi-synonyms, hyperonyms/hyponyms, meronyms/holonyms and other lexical-conceptual relations). But more importantly: this search functionality of the system can use these same relations to show results using not only the words entered by the user, but also these lexical-semantic relations just mentioned. We're talking about what we call "ontological research", as we'll see below (§ 4.3).

The present project aims to transform the *Dicionário Aberto* (DA)[111] into an advanced encoding dictionary. The intention is that, using the reverse search capability in the *DA*, users can search for related lexical units (synonyms, quasi-synonyms, hyperonyms, hyponyms, meronyms, holonyms, etc.) from one or more words.

To make this possible, it will be necessary that the search system available to users works not only with the terms input by the user (and their instances in the macrostructure and microstructure), but also with a semantic structure that provides lexical-conceptual relations between the terms introduced and other terms. These relations will be cross, and proximity measures will be calculated, thereby being possible presenting a set of results sorted by relevance.

4. The advanced search in the *Dicionário Aberto*

The *DA* application Web has evolved, incorporating various types of search: search by entry of article, search in parts of lemma (beginning, middle or end), or

[111] The *Dicionário Aberto* is available online for consultation and for automatic extraction of information in http://www.dicionario-aberto.net, but also for local use, open and free. The project began in June 2005, with the transcript of the 1913 edition of the two volumes of the *Novo Diccionário da Língua Portuguesa,* de Cândido de Figueiredo. For more information about the *Dicionário Aberto*, see Simões & Farinha (2011).

reverse search, among other capabilities. At this moment we are working in a semi-automatic system for extracting synonyms, hyperonyms/hyponyms, meronyms/holonyms that serves to explore the lexical-conceptual relations between terms entered in search (Simões, Iriarte & Almeida, 2012).

Like any electronic dictionary, the *DA* has a search tool. The simple search tool allows the user to find words related to the searched lemma (synonyms, hyperonyms or meronyms), as well as to obtain a set of words orthographically similar, useful for when you don't know the exact spelling of a lemma. The *DA* also has a feature that allows the user to "flip through the dictionary", that is, to browse the dictionary entries sequentially, watching the words near in alphabetical order.

More interesting are the advanced search capabilities, which can turn the dictionary into a real onomasiological dictionary and encoding dictionary at the same time. Using the reverse search and ontological search capabilities, users can search for related lexical units (synonyms, quasi-synonyms, hyperonyms, hyponyms, meronyms, holonyms, etc.) and co-occurring words from one or more words. These advanced features, available for registered users, can be very useful tools for linguists and researchers in Natural Language Processing.

4.1. Search in parts of lemma (beginning, middle or end)

The advanced search allows you to search lemma fragments (beginning, middle or end) by selecting, in the interface, the word "Prefix", "Infix" or "Suffix". These terms, that we hope will be substituted, are not the most appropriate, because the results may not correspond to these morphological categories. Initially, however, we found it clear enough for the general public, and, especially, short enough to be used comfortably in the interface.

We may find, for example, words ending in *–dade* (*abundidade; aceitabilidade; acessibilidade; aceitabilidade; acerbidade; acessibilidade; acetosidade; …*), words beginning with *pre–* (*pre…; pre-romano; preá; preadamita; preadivinhar; preagónico; prealegar;…*) or words, in the original spelling of 1913, containing *–mm–* (*accommodação; accommodadamente; accommodadiço; accommodamento; accommodar; accommodatício; accommodável; …*).

This search capability can be an important tool for morphology studies (Millán, 1999). We can imagine, for example, its usefulness for studying the affixes productivity, such as diminutive suffixes (*–inho, –ito –ino*, etc.); the productivity of certain suffixes in scientific terminology (*–ato, –eto –ito)*; the productivity and real synonymy of suffixes like *–dade/ –ção/ –são, –ança/ –ância*; etc.

The dictionary may prove to be an important resource for linguistic research, and for the development of grammars and other dictionaries, since it allows us to download the results of these searches.

Here's an example on the productivity of affixes: Can all the Portuguese adjectives ended with the suffix *–vel* (like *amável*) form adverbs ending in *–mente* (like *amavelmente*)?

From the *DA* you can download lists of adjectives formed with the suffix –vel and lists of adverbs ending in *–mente* (or, better than the suffix, the form *–velmente*):

…	…
agitável	abominavelmente
aglutinável	admiravelmente
agradável	adoravelmente
agradecível	afavelmente
agricultável	affavelmente
ajuntável	agradavelmente
alcançável	amavelmente
alcoolizável	amigavelmente
alheável	amoravelmente
aliável	aprazivelmente
alienável	civelmente
alliável	comendavelmente
alterável	commendavelmente
amável	compativelmente
amigável	compreensivelmente
…	…

The results can be easily aligned, thus answering the first question (Can all the Portuguese adjectives ended with the suffix *–vel* form adverbs ending in *–mente*?), which then can help us verify hypotheses that explain which can and which cannot:

agitável	…
aglutinável	…
agradável	agradavelmente
agradecível	…
agricultável	…
ajuntável	…
alcançável	…
alcoolizável	…
alheável	…
aliável	…
alienável	…
alliável	…
alterável	…
amável	amavelmente
amigável	amigavelmente
…	…

4.2. The reverse search

As we said, the reverse search capability turns the DA into a real onomasiological dictionary and encoding dictionary. The DA can thus help answer questions like, for instance:

- *"O que acontece à água com o frio?"* ('What happens to the water with the cold?'). Selecting reverse search and typing *água* ('water') and *frio* ('cold') we have the following results: *desnevado, fresco, gelo* e *neve;*

- "Quem é o médico dos olhos?" ('Who is the eye doctor?'). Selecting reverse search and typing *médico* ('doctor') and *olhos* ('eyes') we have the following results: *oculista, oftalmiatra, oftalmologista*;

- "What is the word used in Portuguese to mean 'improve the hardness of metal'? Selecting reverse search and typing *endurecer* ('to harden') and *metal* we obtain as a result the entry **temperar** ('to temper').

Work is being done to ensure that research can be done using not only word forms occurring in the microstructure but also the lemmas corresponding to these word forms, using a morphological analyser.

4.3. The ontological search

I believe that what we call "ontological research" is much more interesting and promising. This research is based on an ontology constructed dynamically (and, therefore, fully automatically).

Unlike what happens with the reverse research, in ontological research, some results are entries that don't contain any of the terms entered by the user in the search window. Take, for example, the following results after introducing the term *mamífero* ('mammal') in the ontological search window:

otária[1] — *f.* ; Género de plantas asclepiadáceas. Espécie de foca, de orelhas bem visíveis....

lontra[1] — *f.* ; Pequeno quadrúpede carnívoro, da fam. das martas.*M. Prov. trasm.*; Pescador...

golfinho[1] — *m.* ; Grande peixe marítimo, carnívoro, da fam. dos cetáceos. *Heráld.*; Represen...

macoco[1] — *m.* ; Animal do Congo, talvez espécie de antílope.

capiguará[1] — *m. Bras*; Espécie de lontra. (Do guar.)

...

We don't find the term *mamífero* ('mammal') in these definitions.

In extraction of lexical-conceptual relations (Simões, Iriarte & Almeida, 2012), some structures present in the definitions were used. Here are some exemples:

– *o mesmo (ou melhor) que* ... ['the same (or better) than'] > Synonymy (SYN);

– *que não é* ... ['that is not'] > Antonymy (ANT);

– *espécie de* ... ['kind of'] > Hyponymy (HIPO);

– *que tem por tipo*... ['which are the type'] > Hyperonymy (HIPER);

– *cada uma das partes que formam* ... ['each of the parties that form'] > Meronymy (MERO);

– *composto por*... ['compound of']> Holonymy (HOLO).

Thanks to the regularity of the structure of the definitions, we can establish a set of rules or patterns (Hearst 1992) using sequences of words that can be found in the definitions. There are high chances of the word that follows the pattern having a lexical-conceptual relation with the lemma of the respective definition.

To the ontological search capability we also use calculated relations (for example, using the transitivity of the relation of hypernym). Thus, mathematical rules were

defined for completion of the ontology rules, inferring new relations from the initial relations:

– a SYN b ⇒ b SYN a (symmetry property between synonyms: if *a* is a synonym for *b*, then *b* is also synonymous for *a*). This relationship is very productive because, in many situations of true synonyms, lexicographers collected the relation of synonymy only in one of the entries of the words involved. Thus, this information can be retrieved for other entry.

– a HIPO b ∧ c HIPO b ⇒ a COHIPO c (If two words, *a* and *c,* are hyponyms of the same word, *b*, then they, *a* and *c,* are co-hyponyms). Relation of co-hyponymy can be calculated from relations of hyponymy.

– a HIPO b ∧ b HIPO c ⇒ a HIPO c. Transitivity of the hierarchical relationships, such as the hypernym or hyponymy, allows one to search for generic terms using a very specific term, or search for specific terms using very generic terms. Think, for example, in a search using the term *animal* ('animal'). It is unlikely that you will find entries for animals. But, by transitivity, we can find other classes as *mamífero* ('mammal') or *peixe* ('fish'), which are hyponyms of "animal". We can find hyponyms terms of *mamífero* ('mammal') or *peixe* ('fish') that are also hyponyms of *animal* ('animal').

We are aware that some rules can be problematic (for example, the case of synonyms and quasi-synonyms). In any case, we believe that a set of possible false synonyms is preferable to no results or drastically reduce the number of ontological relations resulting. Ontology becomes more useful when there is great diversity in the existing relations.

5. Conclusions

Running experiments and implementing algorithms on the DA has been very rewarding. We are convinced the DA will be an excellent tool that cannot only be used as a traditional dictionary, but also as a resource for tasks of natural language processing, as a tool to assist in hypothesis testing for linguistic research and assist in the development of grammars and other dictionaries.

References

Aulete, F. J. Caldas (1987). *Dicionário da Língua Portuguesa Caldas Aulete.* Rio de Janeiro: Editora Delta. [5a edição brasileira, revista, actualizada e aumentada por Hamílcar de Garcia e Antenor Nascentes].

Béjoint, H. (2004). *Modern lexicography: an introduction.* Oxford: University Press.

Casteleiro, J. Malaca (coord.) (2001). *Dicionário da Língua Portuguesa Contemporânea da Academia das Ciências de Lisboa.* Lisboa: Academia das Ciências de Lisboa/Editorial Verbo.

Ferreira, A. Buarque de Holanda (1999). *Novo Aurélio Século XXI: O Dicionário da Língua Portuguesa.* Rio de Janeiro: Nova Fronteira.

Figueiredo, C. de (1982). *Grande Dicionário da Língua Portuguesa.* Lisboa: Livraria Bertrand.

Haensch, G. (1982). "Tipología de las obras lexicográficas". In Haensch, G.; Wolf, L.; Ettinger, S.; Werner, R. (eds.). *La lexicografía. De la linguística teórica a la lexicografía práctica.* Madrid: Gredos. 95-187.

Hearst, M. (1992). "Automatic acquisition of hyponyms from large text corpora". *Proceedings of the Fourteenth International Conference on Computational Linguistics,* Nantes, France. Volume 2. 539–545.

Houaiss, A. (2001). *Dicionário Houaiss da Língua Portuguesa.* Rio de Janeiro: Objectiva.

Martínez de Sousa, J. (1995). *Diccionario de lexicografía práctica.* Barcelona: Bibliograf.

Millán, J. A. (1999). "Zigzag, gong, ping-pong, iceberg. donde se descubre que hay diccionarios inversos, y su utilidad manifiesta para el progreso de la humanidade" [on-line]. http://jamillan.com/inverso.htm [Access date: 1 July. 2013]

Ramalho, E. (1985) *Dicionário Estrutural, Estilístico e Sintáctico da Língua Portuguesa.* Porto: Lello & Irmão Editores.

Simões, A.; Farinha, R. (2011). "Dicionário Aberto: Um novo recurso para PLN", *Vice-versa* 16, 159–171.

Simões, A.; Iriarte, Á; Almeida, J. J. (2012). "Dicionário-aberto – a source of resources for the portuguese language processing". In Caseli, H.; Villavicencio, A.; Teixeira, A.; Perdigão, F. (eds.) *Computational Processing of the Portuguese Language, Lecture Notes for Artificial Intelligence,* Berlim: Springer, 7243, 121–127.

Simões, G. A. (1994). *Dicionário de Expressões Populares Portuguesas.* Lisboa: Dom Quixote.

Vilela, M. (1991). *Dicionário do Português Básico.* Porto: Edições Asa.